Pilot-Study Proposal

Total word count: 724

1. Type of predictive task

The given predictive task tells us that we should determine whether the customer would file a claim further or not. Therefore, knowing that we can have only two different possible results such as “Yes” or “No”, we can apply a certain classification ML technique to classify the customer according to his claim history and historical data of past policies.

Besides that, applying the rules mining method could fit perfectly to this sort of problem, because it may reveal in the data even indistinct rules that could better describe a financial state of a customer. Along with that, it may point out what are the potential risks concerning the insured one. Thus, the manager can take that information into account to make a proper decision regarding a discount.

Considering the opportunities of modern ML classification models, this type of predictive task can be successfully solved by some of them only relying on retrieved features from the mentioned set of data.

2. Examples of possibly informative features

First of all, while considering a discounted premium suggestion, the insurance manager would like to know what is the bank balance of the client. For example, if in the past the insured one travelled many times, but now he has some financial difficulties, it would be a good sign that he may obtain a discounted premium for the next trips.

Also, the most important features are purpose, destination, and time of the travel, because they directly influence the amount of the premium and the chance to obtain it. As an example, if a client has a professional sports activity, he is likely to obtain a higher discount on the next journey, as his life is often under threat.

Besides, the age of the insured one potentially could be a good predictor. The reason for it is the probability of having some severe diseases such as cancer or diabetes which can affect the traveller suddenly. Apparently, in the event of having similar diseases, a client will necessarily file a future claim on future travel.

3. The proposed learning procedures

According to the description of the given predictive task, we should consider the following ML techniques which can successfully solve the problem:

- Decision trees

  Since the decision trees method is based on rules mining it should be rather efficient while processing an entire dataset of customers' data. Thus, we shall obtain a comprehensive decision tree that can be applied to the classification of a new insured customer.

- K-nearest neighbours

- This learning procedure can be easily applied to this task, because of its relatively simple approach. As the result, this classifier will seek k-most similar cases to a given insured customer in the historical data. So, if we had a sufficiently large dataset, the "K-nearest neighbours" method would have a decent classification accuracy.

- Support vector machines

  Since support vector machines is a rather powerful mathematical instrument of classification, it should provide us with high accuracy after being applied to the mentioned dataset. Therefore, it is expected to succeed in the determination of a new customer's type.

4. Performance evaluation

To evaluate the performance of our selected classification methods, besides the simple classification accuracy, we can also use such metrics as precision, recall, and F1 score. The reason for it: they can better describe the model's accuracy on a dataset with unequal amounts of different classes. As the result, we should rely on an F1 score to understand the real accuracy.

However, the performance evaluation procedure, in general, is conducted in two stages: classification accuracy calculation on a training dataset and evaluation of those metrics on test data.

By the way, we can implement a cross-validation technique for the first stage thereby making our model's estimation more precise. This approach should prepare the trained model for the final test.

Furthermore, we can compare the performance of several classifiers using bar charts for a clear demonstration. Also, for example, we can reflect on a plot a time that was taken for training. So it can be seen what was the fastest algorithm and, respectively, the slowest one.

All in all, performance evaluation often is a rather complicated task, because even after multiple tests of our model we still cannot know for sure that it would always perform as planned.